Available online at www.sciencedirect.com

ScienceDirect

Genomics 91 (2008) 335–346

GENOMICS

www.elsevier.com/locate/ygeno

Evolution in the laboratory: The genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1 ☆

F. Pfeiffer [a], S.C. Schuster [a,1], A. Broicher [a], M. Falb [a,2], P. Palm [a], K. Rodewald [a], A. Ruepp [b,3], J. Soppa [c], J. Tittor [a], D. Oesterhelt [a,*]

[a] *Department of Membrane Biochemistry, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany*
[b] *Department of Molecular Structural Biology, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany*
[c] *Institute for Molecular Biosciences, Goethe University, Frankfurt, Germany*

Received 2 August 2007; accepted 2 January 2008
Available online 3 March 2008

Abstract

We report the sequence of the *Halobacterium salinarum* strain R1 chromosome and its four megaplasmids. Our set of protein-coding genes is supported by extensive proteomic and sequence homology data. The structures of the plasmids, which show three large-scale duplications (adding up to 100 kb), were unequivocally confirmed by cosmid analysis. The chromosome of strain R1 is completely colinear and virtually identical to that of strain NRC-1. Correlation of the plasmid sequences revealed 210 kb of sequence that occurs only in strain R1. The remaining 350 kb shows virtual sequence identity in the two strains. Nevertheless, the number and overall structure of the plasmids are largely incompatible. Also, 20% of the protein sequences differ despite the near identity at the DNA sequence level. Finally, we report genome-wide mobility data for insertion sequences from which we conclude that strains R1 and NRC-1 originate from the same natural isolate. This exemplifies evolution in the laboratory.

Keywords: Archaea; Comparative genomics; Genome sequence; Halophilicity; *Halobacterium salinarum;* Insertion sequence

Halobacterium salinarum has been intensively studied during the past decades, through which our understanding of various biological processes such as energy metabolism, environmental response, gene regulation, and the archaeal cell cycle has been greatly increased (for recent reviews see [1,2]). A microorganism corresponding to the description of *Hbt. salinarum* was isolated from salted fish more than 80 years ago. Since then many haloarchaeal species have been isolated, which, after considerable renaming, are currently grouped into 25 genera. Several years ago, it was decided that the species *Halobacterium salinarum, Halobacterium halobium,* and *Halobacterium cutirubrum* are so similar that they should be regarded as strains of one species named *Halobacterium salinarum* [3]. *Hbt. salinarum* shows very high genetic variability [4,5] that was attributed to the large number of insertion sequences (ISH elements) (for review see [6]).

The active and successful research of several laboratories has led to the initiation of two independent genome sequencing initiatives, one for *Halobacterium* sp. NRC-1, the other for *Hbt. salinarum* strain R1. When the sequence of *Hbt.* sp. NRC-1 genome appeared [7], the chromosome of *Hbt. salinarum* strain R1, which is reported here, was complete and being annotated. The assembly of the smaller replicons had not been finished at that time due to major problems caused by large-scale duplications and the high number of insertion elements.

☆ Sequence data from this article have been deposited with the DDBJ/EMBL/GenBank Data Libraries under Accession Nos. AM774415 (chromosome), AM774416 (pHS1), AM774417 (pHS2), AM774418 (pHS3), and AM774419 (pHS4).

* Corresponding author. Fax: +49 89 8578 3557.
E-mail address: oesterhe@biochem.mpg.de (D. Oesterhelt).
[1] Current address: Center for Comparative Genomics and Bioinformatics, Center for Infectious Disease Dynamics, Penn State University, University Park, PA 16802, USA.
[2] Current address: Sanofi-Aventis Deutschland GmbH, Industriepark Hoechst, 65926 Frankfurt am Main, Germany.
[3] Current address: Institut für Bioinformatik, GSF-Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany.

0888-7543/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.01.001

Genome sequences from different strains of the same organism may vary significantly; for example, there is a 5–7% sequence deviation between the two *Helicobacter pylori* strains J99 and 26695 at the amino acid level [8]. In contrast, the number of sequence differences for *Hbt. salinarum* strains R1 and NRC-1 was vanishingly small, although NRC-1 had been published as if it were a distinct species. However, since then it has been reclassified as a strain of *Hbt. salinarum* [9]. It was also evident that the protein-coding gene sets differed considerably, although they were derived from nearly identical DNA sequences. This inconsistency is due to the high error rate of automatic gene finder programs for GC-rich genomes, especially with respect to start codon selection [10–13]. The correctness of the protein-coding gene set can be increased by experimental or bioinformatic analysis, for example, integration of proteomic data or evaluation of sequence homology data.

The genome of *Hbt. salinarum* strain R1 has been publicly available since 2002 through the HaloLex Web portal (www.halolex.mpg.de). In this publication, we report the sequences of the chromosome and the four plasmids of *Hbt. salinarum* R1, including a high-quality annotation of protein-coding genes that is well supported by proteomic experiments and sequence homology data. Comparison of strains R1 and NRC-1 revealed genome-scale data on sequence differences and ISH element mobility. From these results, we conclude that both strains originate from the same natural isolate and have since diverged in the laboratory.

Results and discussion

The genome of Halobacterium salinarum strain R1

The genome of *Hbt. salinarum* strain R1 (DSM 671) comprises a single major chromosome of 2 Mb with a very high GC content of 68.0% and four megaplasmids (pHS1 to pHS4) that have a total of 667,814 bp and a lower GC content of 58.8% (Table 1). The genome contains 2878 protein-coding genes (Supplementary Table S1).

Table 1
Basic characteristics of the replicons from *Halobacterium salinarum* strain R1 (DSM 671)

	Chromosome	pHS1	pHS2	pHS3	pHS4
Length (bp)	2,000,962	147,625	194,963	284,332	40,894
GC content	68.0%	57.4%	58.6%	59.8%	57.9%
% coding (protein + RNA)	91.5%	83.6%	80.4%	84.8%	82.9%
Encoded proteins	2,132	172	230	305	39
Average protein length (amino acids)	284	239	226	263	289
Encoded stable RNAs	52	—	—	—	—

	Plasmids (total)	Genome (total)
Length (bp)	667,814	2,668,776
GC content	58.8	—
Encoded proteins	746	2,878

The major chromosome

The chromosome is densely packed with 2132 protein-coding genes and genes for 52 stable RNAs. Together, these cover 91.5% of the chromosomal sequence.

The replication origin is delineated by a 31-bp inverted repeat that is flanked on one side by a Cdc6 homolog (*orc7,* OE4380F) [14]. On the other side the repeat is flanked by a set of three genes (OE4377R, OE4376R, OE4374R) that are also found adjacent to the replication origin in *Natronomonas pharaonis* [10], *Haloquadratum walsbyi* [15], and *Haloarcula marismortui* [16].These genes have no known function, but the positional conservation observed in all halophiles may indicate an involvement of the three proteins in the replication process.

The chromosome contains a 60-kb insertion with plasmid-like characteristics: (a) a reduced GC content of 56% and (b) a reduced proteomic protein identification rate [17]. This insertion corresponds to the previously described "AT-rich island" [18].

Plasmid pHS3

Plasmid pHS3 is 284 kb long and codes for a number of essential and important proteins, most of them in or adjacent to a 67-kb region (Fig. 1A) with chromosome-like features (increased GC content of 65%, increased proteomic identification ratio) [17]. The most prominent proteins are (a) the only arginine-tRNA ligase of *Hbt. salinarum* (argS); (b) the two subunits of aspartate carbamoyltransferase (pyrBI), which catalyzes the first step of pyrimidine biosynthesis; (c) all enzymes of the arginine deiminase pathway for arginine fermentation (arc operon) [19], including the arginine/ornithine antiporter (OE5204R, unpublished data); and (d) the only catalase (perA) which is involved in protection against oxidative stress. Thus, because pHS3 encodes essential proteins it may be considered a second chromosome rather than a plasmid.

Plasmids pHS1, pHS2, and pHS4

The three plasmids pHS1, pHS2, and pHS4 are related to each other through their large-scale duplications (Fig. 1A). Regions that are labeled by the same letter show (near) sequence identity. The regions are listed in Supplementary Table S2.

The 147-kb plasmid pHS1 corresponds to the previously described plasmid pHH1 [20] and carries a high number of ISH elements. Only one-third (48 kb) of the pHS1 sequence is specific for this plasmid (regions B, G, L, M). The other two-thirds (99 kb) represent three large-scale duplications (Fig. 1A): a perfect 61.8-kb duplication of pHS2 (regions C, D, F), a perfect 30.0-kb duplication of pHS4 (region K) adjacent to an imperfect duplication with 98.5% sequence identity over 7.3 kb (region H).

Plasmid pHS2 is 195 kb long, of which 61.8 kb are duplicated on pHS1 and the remaining 133 kb are specific to pHS2.

Plasmid pHS4 with 41 kb was not detected until the late stages of genome assembly since 92% of it represents sequences duplicated on pHS1, while only 8% of the sequence (3.4 kb, region Y) is specific to pHS4. There is a perfect 30.0-kb duplication (region K) and an adjacent imperfect duplication of 7316 matching bases with only 1.5% sequence difference (region H). An additional difference is the presence of two ISH elements, which occur only on pHS1 in region H.

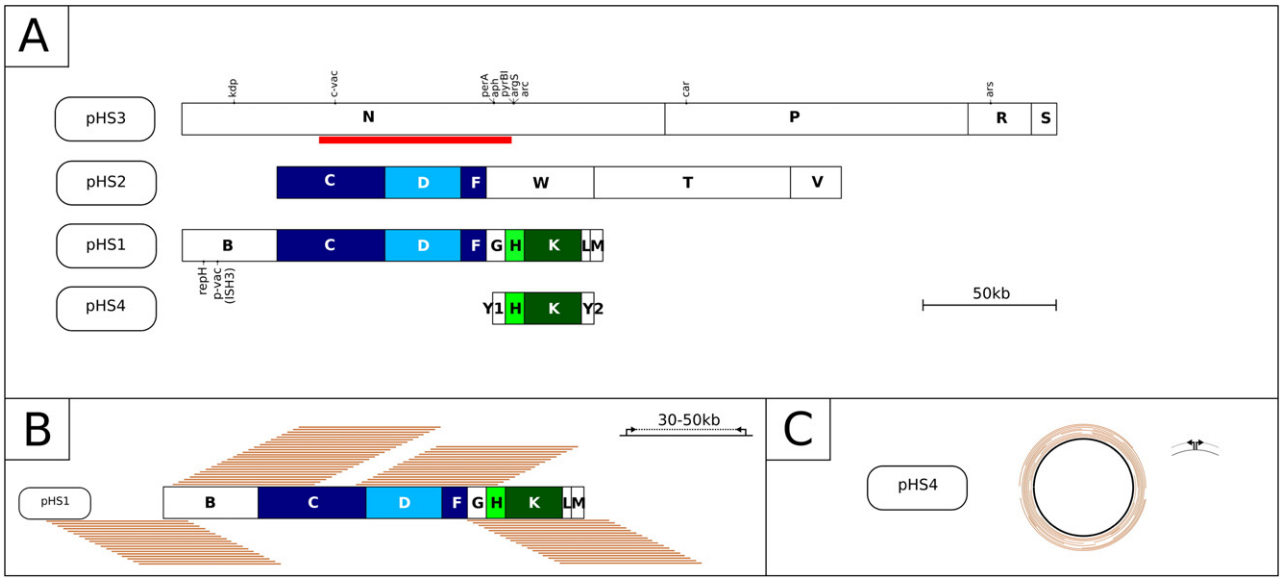

Fig. 1. Schematic representation of the plasmids from *Hbt. salinarum* strain R1 and their validation. (A) Structures of the plasmids from strains R1. Plasmids pHS1, pHS2, pHS3, and pHS4 from strain R1 are schematically represented as linearized scaled bars. The regions are labeled with letters, subregions with an additional number. Plasmid names are shown at the left. Duplications of pHS1 on pHS2 are drawn in blue, those on pHS4 in green. Nonduplication regions are drawn in white. Scaling of the regions is based on their length. For graphical reasons, short regions (<2 kb) are shown slightly oversized. The location of the 67-kb GC-rich region in pHS3 is indicated by the red bar. Several relevant genes are marked above pHS3 or below pHS1. All plasmids are circular. For all plasmids, the base numbering begins at the left end, except for pHS2, for which base 1 is at the beginning of region T. (B) Cosmid validation of pHS1. Plasmid pHS1 is shown linearized with an indication of all cosmids (brown lines) that validate its structure. The 18 cosmids traversing the circularization point are indicated at both cosmid ends below the plasmid. Cosmids not traversing the circularization point are indicated above the plasmid. The basic technique is outlined in the top left corner: Cosmid end sequences must be oriented toward each other and must be 30–50 kb apart. (C) Cosmid validation of pHS4. Plasmid pHS4 is represented by the central black circle. Cosmids are indicated by brown open circles according to the position of the terminal sequences. Cosmid ends are evenly distributed all over pHS4. The basic technique is outlined in the top right corner: As the cosmids represent (nearly) all of the plasmid, end sequences are close to each other on the plasmid sequence but point in opposite directions.

The origin of replication has been identified for pNRC100 [21] and for pHH1 [20], both of which are closely related to pHS1. It is located between the *repH* gene (OE7014F, a plasmid replication protein) and the preceding divergently transcribed gene for OE7012R. Two other rep genes (*repI, repJ*) are located in regions H and K, which are both duplicated on pHS4 and thus may be involved in replication of pHS4. However, plasmid pHS2 does not contain a rep gene. It does contain several Cdc6 family proteins. It should be noted that the chromosomal replication protein Orc7 is also a member of this protein family. Plasmid pHS3 also lacks a rep homolog but contains several cdc6 homologs. A similar situation is found in *Har. marismortui* [16], in which only two of seven plasmids code for a rep gene, while the other five code for at least one Cdc6 family protein.

Validation of the plasmid assembly

Repetitive sequences such as large sequence duplications or ISH elements severely interfere with genome assembly in general. To overcome this problem, additional methods are required to delineate the correct connectivity of the sequences adjacent to the repeat. Cosmids with their average length of 40 kb and method-inherent lower and upper size limits (30–50 kb) are the method of choice.

Cosmid end sequences were determined and positioned onto the assembled plasmids. A dense set of suitable cosmids was obtained in which the cosmid end sequences show convergent orientation as well as method-coherent distance. This set of suitable cosmids provides a dense scaffolding through which the structure of all plasmids has been unequivocally confirmed as illustrated for plasmid pHS1 (Fig. 1B). As an example, the circularization point between regions M and B was validated with 18 distinct cosmids.

The perfect 61.8-kb duplication common among plasmids pHS1 and pHS2 even exceeds the maximum cosmid length. It is still unclear how such a long duplication, which shows an identical sequence, can exist in an organism capable of homologous recombination. An experiment was designed to validate the colinearity of the two plasmids over the whole length of the duplication. This experiment is based on the analysis of cosmids that extend across the boundary between plasmid-specific and duplicated regions. On these cosmids, a nested set of overlapping 7-kb PCR fragments resulted in corresponding products within the duplication and plasmid-specific products across the boundary (Supplementary Fig. S3).

Cosmid data also confirm the existence of plasmid pHS4. A set of 29 cosmids originate from cloning of all (or the majority) of pHS4 (Fig. 1C). In this case, cosmid ends show divergent orientation, being positioned close to one another on the assembled pHS4 sequence. Cosmids with this configuration are evenly distributed over the whole plasmid.

A reliable set of protein-coding genes for Halobacterium salinarum

It is well established that gene prediction is difficult in GC-rich genomes [10,12,13]. Severe ORF overprediction results in

two types of problems: (a) The existence of long alternate open reading frames [22,23] makes it difficult to discriminate protein-coding genes from spurious ORFs. (b) Start codon selection is highly error-prone due to long N-terminal ORF extensions in front of the start codon used in vivo. These extensions, which reflect the large distance to the nearest preceding in-frame stop codon, may contain several alternative start codons [11,13].

We used several approaches to achieve a high-quality set of protein-coding genes. (i) An extensive set of proteomic data has been collected, which led to the identification of 1958 proteins [11,17,24–27], thus validating the assigned reading frame. For 606 proteins, the N-terminal peptide could be reliably identified, hence unambiguously validating the assigned start codon [11,25]. (ii) We applied intergenomic comparison to three other halophiles (*Nmn. pharaonis* [10], *Har. marismortui* [16], and *Hqr. walsbyi* [15]). Many of the proteins are well conserved within the true coding region, while putative but spurious N-terminal extensions are devoid of sequence similarity. (iii) Other characteristics were also applied, such as the acidic p*I* value of halophilic proteins which differs from highly basic p*I* values commonly found in spurious ORFs.

The resulting set contains 2878 protein-coding genes for *Hbt. salinarum* strain R1 of which 68% have been identified by proteomics (indicated in Supplementary Table S1). More than 100 identified orphan proteins (proteins with unknown function and without homologs) are included in this set. Apart from the protein-coding genes, there are also 6517 spurious ORFs, 96% of which are longer than 100 codons, the longest having 1341 codons. Despite the extensive set of genome-wide proteomic data, the usage of alternate overlapping reading frames has not been detected upon stringent analysis of our proteomic data and thus, if such multiple usage occurs at all, it must be a very rare event. The same result was obtained for *Nmn. pharaonis* [22].

It should be stressed that a high-quality gene set is fundamental to other research areas. Genetic experiments (for example, gene deletion, protein overexpression) critically depend on a correctly assigned start codon. Also, leaderless transcripts can be detected only for genes with correctly annotated start codons. If the annotated ORF is too long, probes for transcriptomic studies may overlap with the neighboring gene and lead to invalid results. Also, identification of N-terminal peptides by proteomics would be hampered if the start codon is misassigned. The prediction of protein export signals such as signal sequences and twin-arginine motifs is commonly restricted to the N-terminal regions and thus is affected by erroneous start codon selection.

Comparison of the genomes from strains R1 and NRC-1

Comparison of the chromosomes

The comparison of strain R1 (this report) with strain NRC-1 [7], published as *Halobacterium* sp. NRC-1, revealed complete colinearity of the chromosome and nearly identical DNA sequences. Aside from differences related to ISH elements, there are only 12 other differences: four point mutations, five single-base frameshifts, and three insertion/deletion events (Table 2). For the majority of the described differences, additional validation of the R1 sequence by experimental proteomic data is available.

Three of the four single-base exchanges cause amino acid substitutions, the fourth is silent (Table 2). Three of the five single-base frameshifts have a major effect on the protein sequence. One is within a coding sequence that additionally contains two strain-specific ISH elements. The last single-base "frameshift" is located in an intergenic region.

The three insertion/deletion events merit a more detailed description.

Table 2
The 12 differences between the chromosomes of strains R1 and NRC-1

Type	Position	Base(s) (R1 → NRC-1)	R1	NRC-1	Description
Base change	5697	C → G	OE1013R	VNG0006G	Silent mutation
Insertion/ deletion 1	175135–175142	—			10,007-bp insert in NRC-1; 8-bp target duplication
Base change	350453	A → C	OE1695R	VNG0466C	Point mutation Ser-22 → Ala; Ser-22 validated by proteomics
Frameshift	416705	C → CC	OE1823F	VNG0553C	Divergent beyond position 392; acidic p*I* only for R1 sequence
Frameshift	452158	G → GG	OE1916F	VNG0606G	Start codon out of frame in NRC-1; VNG0606G starts with Met-53; peptides before Met-53 validated by proteomics
Frameshift	578093	C → CC	OE2141F	VNG0779C + VNG0780H	Frameshift at pos 528 in VNG0779C; C-terminus starting with Met-573 is VNG0780H; peptides after pos 528 validated by proteomics
Frameshift	628043	A → —	—	—	This "frameshift" is located in an intergenic region
Base change	665004	A → C	OE2303F	VNG0887G	Point mutation Lys-544 → Asn; Lys-544 validated by proteomics
Base change	1016811	C → A	OE2961F	VNG1347G	Point mutation Arg-208 → Ser
Frameshift	1221555	C → —	OE3338R	VNG1650H	In addition to frameshift, VNG1650H is interrupted by two insertion sequences
Insertion/ deletion 2	1615347–1615766	—	OE4073R	VNG2196G	NRC-1 lacks second halocyanin domain; R1-specific peptides validated by proteomics
Insertion/ deletion 3	1863363	—	—	—	133 additional bases in NRC-1 affecting rRNA promoter region

The differences between the chromosomes from strains R1 and NRC-1, excluding those related to ISH elements, are presented. For single-base changes and one-base frameshifts, the differences at the DNA and protein levels are indicated. Also, the positions and some details of the three insertion/deletion events are shown. Correctness of the R1 sequence at the difference points was confirmed by counterchecking with the raw sequencing data. Additional evidence on the protein level, which supports the R1 sequence, is specified as "validated by proteomics" and indicates reliable identification by tandem mass spectrometry.

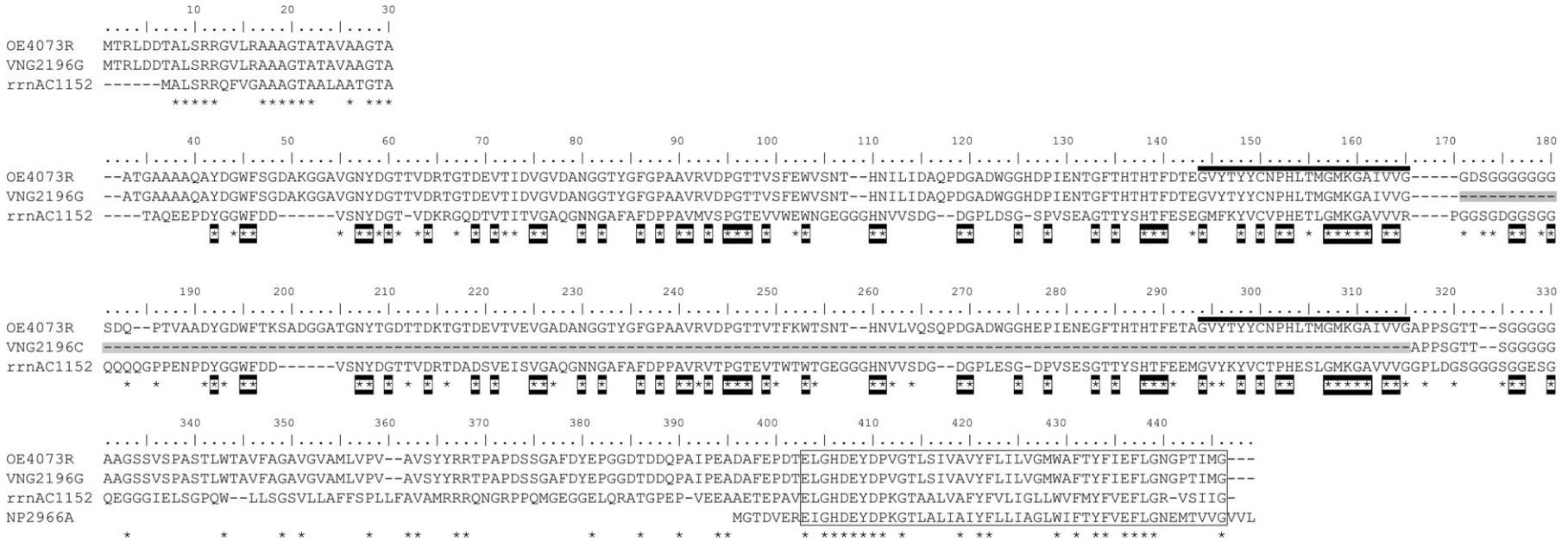

Fig. 2. Multiple alignment of halocyanin *hcpB*. Multiple alignment of *hcpB* from *Hbt. salinarum* (strain R1, OE4073R; strain NRC-1, VNG2196G) and *Har. marismortui* (rrnAC1152) with the *cbaD* gene of *Nmn. pharaonis* (NP2966A). The two central blocks represent the copper-binding domains (alignment positions 41–165 and 191–315). The second copper-binding domain is missing in strain NRC-1 (indicated by the gray line). Amino acids that are identical in all aligned sequences are indicated by asterisks. Asterisks are boxed when residues occur in both copper-binding domains of both organisms. The region of 22 consecutive identical amino acids in the two domains of OE4073R is indicated by a bar above the sequence. The *cbaD* domain occurs at the extreme C-terminus (boxed, alignment positions 403–446).

Insertion/deletion 1. Strain NRC-1 contains an insertion of 10,007 bp compared to strain R1. This region is flanked by an 8-bp target duplication, indicating that the insert originates from a transposition event in strain NRC-1.

Insertion/deletion 2. Compared to strain R1, strain NRC-1 contains an in-frame 423-bp deletion within the halocyanin gene *hcpB* (OE4073R/VNG2196G). The carboxyl-terminal region of *hcpB,* which is present in both strains as well as in *Har. marismortui,* is homologous to the 9-kDa subunit *cbaD* of the cytochrome-*c*-type terminal oxidase from *Nmn. pharaonis* [10,28]. In *Halobacterium* and *Haloarcula,* this small subunit of terminal oxidase has been fused to copper-binding halocyanin domains, confirming the previous assumption that copper-containing halocyanins rather than iron-containing cytochromes are involved in electron transfer to the terminal oxidase in the respiratory chain of halophilic archaea [28].

The *hcpB* gene in strain R1 encodes two copper-binding domains that exhibit 76% sequence identity (Fig. 2). *hcpB* from *Har. marismortui* has an identical domain architecture. In contrast, *hcpB* from strain NRC-1 carries an in-frame 423-bp deletion that results in the complete and perfect elimination of the second copper-binding domain so that the protein may remain functional. The domain elimination was probably caused by homologous recombination in a stretch of 32 identical bases (AACCCCCATCTCACGATGGGGATGAAAGGCGC), coding for a region of 22 consecutive identical amino acids between the two domains in R1 (Fig. 2).

Insertion/deletion 3. Strain NRC-1 contains an additional sequence of 133 bp in the promoter region of the rRNA operon. This additional sequence contains a repeated stretch of 27 bp, which occurs twice in strain R1 and three times in strain NRC-1. This repeated sequence has been implicated in rRNA transcription [29] and thus the difference may affect the strength of the rRNA promoter.

Overall, such a vanishingly small number of sequence differences demonstrates the extremely close relationship between the two strains as well as the high fidelity of both genome sequences.

Comparison of the plasmids

The plasmids from strain R1, which are structurally confirmed by a dense set of cosmids (Figs. 1B and C), are compared to the plasmid structures reported for strain NRC-1 [7,30] in Fig. 3.

It is possible to match more than 350 kb of plasmid sequence among the two strains. The matched regions are virtually identical at the DNA sequence level, with just one single-base change and one hot spot of sequence differences. The plasmids contain additional sequences that cannot be matched. These unmatched regions are restricted to 4.5 kb in strain NRC-1 but amount to 210 kb in strain R1.

The similarity at the DNA sequence level is contrasted sharply by a highly different overall plasmid architecture, which is evident at three levels. First, the number of plasmids is different. Nearly all of the sequence from the two plasmids of strain NRC-1 can be matched to the four plasmids from strain R1. Second, large-scale duplications are reported for both strains but the duplication patterns are highly dissimilar. Third, regions of colinearity are comparably short so that colinearity breakpoints are frequent. In addition, all colinearity breakpoints are associated with ISH elements.

The differences in overall plasmid architecture may reflect biological variation among the strains. Alternatively, the excessive duplications may have resulted in sequence assembly errors.

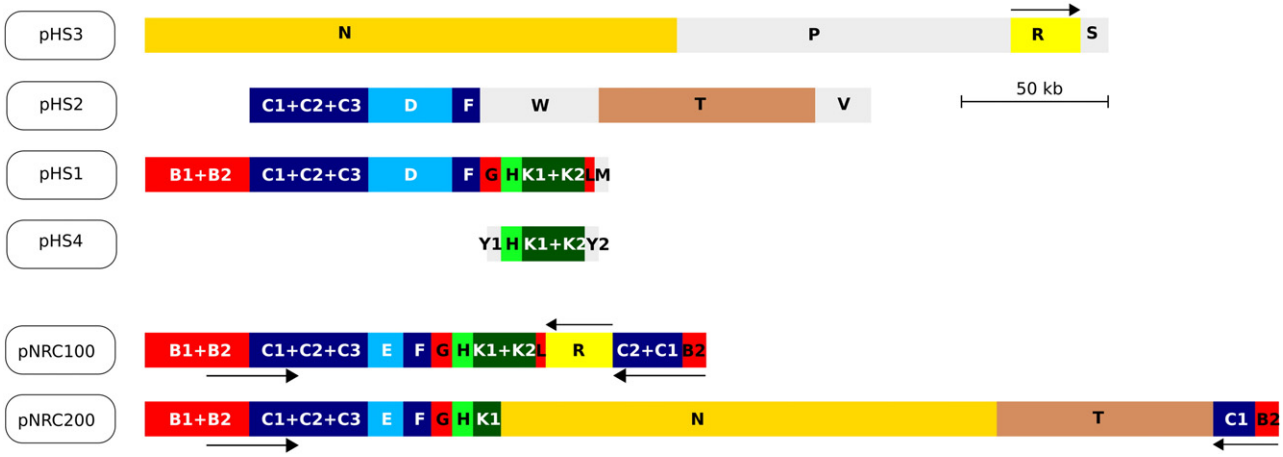

Fig. 3. Schematic representation of the plasmids from *Hbt. salinarum* strains R1 and NRC-1 and their comparison. The architecture of the plasmids from strains R1 and NRC-1 and their matching regions are illustrated. Plasmids pHS1, pHS2, pHS3, and pHS4 from strain R1, as well as pNRC100 and pNRC200 from strain NRC-1, are schematically represented as linearized scaled bars. The regions are labeled with letters, subregions with an additional number. Plasmid names are shown at the left. Duplications of pHS1 with pHS2 are drawn in blue and those with pHS4 are drawn in green. Regions that occur only in strain R1 are drawn in gray. Other regions that match among the plasmids of strains R1 and NRC-1 are indicated in red for pHS1/pNRC100 (partially also present on pNRC200), bright yellow for pHS3/pNRC100, dark yellow for pHS3/pNRC200, and brown for pHS2/pNRC200. The inverted duplications in pNRC100 and pNRC200 are indicated by forward and reverse arrows, respectively. All regions are oriented identically except for the inverted duplications and for region R (also indicated by arrows). Scaling of the regions is based on their length in strain R1 and may differ slightly from that in strain NRC-1 due to strain-specific ISH elements. For graphical reasons, short regions (<2 kb) are shown slightly oversized. All plasmids are circular. For all plasmids, the base numbering begins at the left end, except for pHS2, in which base 1 is at the beginning of region T.

Assembly errors might better explain why virtually identical DNA sequences are found in plasmids that are highly different in overall architecture.

The patchwork of matching regions and duplications in the plasmids is illustrated in Fig. 3. The regions of colinearity and sequence identity are highlighted with letters and color coding. Breakpoints between regions are commonly associated with ISH elements. Full details for all regions and for the breakpoint-associated connecting ISH elements are provided in Supplementary Table S2 and Supplementary Text S4.

Plasmids pHS1 and pNRC100 are colinear for a total of 127 kb (regions B, C, and F, G, H, K, L) (Fig. 3). Colinearity is interrupted between regions C and F by the presence of strain-specific alternative sequences (the 19.3-kb region D in pHS1 and the 4.5-kb region E in pNRC100). The colinear region contains one copy of each of the large-scale duplications. On one hand, duplications occur among the plasmids from strain R1 (regions C+D+F, K, and H). On the other hand, duplications occur among the plasmids from strain NRC-1 (112-kb duplication of regions B+C+E+F+G+H+K1, with regions B2, C1, and C2 also forming the inverted repeats).

Beyond the end of the colinear region, plasmids pHS1 and pNRC100 diverge completely (Fig. 3). In pHS1, the end of region L is 1.9 kb from the circularization point. It is emphasized that circularization at this point is validated by a dense set of cosmids (Fig. 1B). In contrast, region L of pNRC100 is connected to the 16-kb region R. Region R is further connected to an inverted duplication of 40 kb (regions B2, C1, and C2). Both connections are associated with ISH elements. The circularization point of pNRC100 is located at the end of the inverted duplication, a connection that, again, is associated with an ISH element. This architecture described for pNRC100 can be excluded for pHS1 from strain R1, as detailed analysis of cosmid data did not reveal any evidence for an inverted duplication. In addition, region R is found on plasmid pHS3, a placement that is also validated by cosmids.

The first 112 kb of pNRC200 are reported to be identical to that in pNRC100 (regions B, C, E, F, G, H, and part of region K) [7] (Fig. 3). This covers most of the three large-scale duplications among the plasmids from strain R1: the perfect 61.8-kb duplication among pHS1 and pHS2 (regions C, D, and F), the perfect 30.0-kb duplication among pHS1 and pHS4 (region K), as well as the imperfect 7.3-kb duplication among pHS1 and pHS4 (region H). Plasmid pNRC200 also contains the strain-specific alternative 4.5-kb region E of pNRC100. At the point within region K where pNRC100 and pNRC200 diverge, pNRC200 contains an ISH element that is present neither in pNRC100 nor in the plasmids from strain R1. Plasmid pNRC200 continues with the 148.8-kb region N, followed by the 63.1-kb region T, with an ISH element between the two regions. This part of pNRC200 matches to two independent plasmids in strain R1. Region N is part of pHS3, while region T is part of pHS2. In pNRC200, region T is adjacent (with yet another connecting ISH element) to a slightly shorter form of the 40-kb inverted duplication already described for pNRC100 (restricted to regions B2 and C1) [7]. The circularization point of pNRC200 corresponds to that of pNRC100.

It should be noted that connecting ISH elements are associated with all colinearity breakpoints that occur between strains NRC-1 and R1. In our hands, ISH elements regularly caused misassemblies by the genome assembly program Phrap. Therefore, we consider it quite likely that the reported plasmid architecture of pNRC100 and pNRC200 is incorrect to some extent.

The plasmids from strain R1 contain 210 kb of sequences that do not occur in strain NRC-1 (regions W and V on pHS2, P and S on pHS3, Y1/Y2 on pHS4, and M on pHS1). Overall, these 210 kb code for typical haloarchaeal proteins, predominantly (conserved) hypothetical proteins. Region V codes for one of the transcription factor B (TFB) homologs. Region P contains the *car* gene for the sensory transducer mediating arginine chemotaxis [31]. This region also codes for other genes with various functions, for example, ABC-type transport proteins, helicase homologs, or an AAA-type ATPase.

The only hot spot of DNA sequence variation among the plasmids from strains R1 and NRC-1 is located in region H, which shows a 1.5% sequence difference between pHS1 and pHS4 (89 base changes). In this region, pNRC100/pNRC200 differ by 12 bases from pHS1, 9 of which match those from pHS4 (for details, see Supplementary Text S4). This may indicate that a corresponding imperfect duplication also exists in strain NRC-1 but has not been recognized yet.

Comparison of the protein-coding gene sets

Despite the near identity of the DNA sequences of strains R1 and NRC-1, there are major differences in the protein-coding gene set. In our analysis, there are 111 protein-coding genes that have not been annotated for strain NRC-1. Of these, 26 have been confirmed by proteomics. In addition, 47 spurious ORFs are annotated as genes in NRC-1, but these do not belong to the set of protein-coding genes according to our analysis (Supplementary Table S5). A total of 2375 protein-coding genes map to each other in the two strains. Only 1900 protein sequences are identical, while 475 (20%) differ. Taken together, this illustrates the severe ORF prediction problem in GC-rich genomes and emphasizes the necessity to achieve a high-quality gene set. Most of the sequence differences (449 genes) are caused by selection of different start codons (as listed in Supplementary Table S6). For a large number of the genes, a clear decision between alternative start codons is possible based on either the identification of the N-terminal peptide by proteomics (91 genes) or unambiguous results from sequence homology analysis with other haloarchaeal strains (85 additional genes when only homologs from *Nmn. pharaonis* and *Hqr. walsbyi* are counted). Overall, such unambiguous evidence is available for 176 of the 449 genes (40%) with alternative start codon selection and in all cases the available data support the start codon assignments made for strain R1. For the remaining 60% of the genes, additional but weaker evidence supports the R1 start codon assignments (for example due to resolution of gene overlaps or strong p*I* value shifts in the vicinity of the correct start codon [17]. Additional evidence from sequence homology is available for proteins that do not occur in *Nmn. pharaonis* or *Hqr. walsbyi*. Also, homology data may not meet the stringent criteria to be considered unambiguous.

Insertion sequences (ISH elements)

Nearly 100 ISH elements were identified in the genome of *Hbt. salinarum*. These elements account for most of the genetic variability of the organism [4]. Genome-wide ISH element analysis was performed by detailed comparison of the location patterns in the chromosome and the plasmids of strains R1 and NRC-1. Data are reported for "canonical" ISH elements (i.e., those types of ISH elements listed in [6]). When counting ISH elements that occur in large-scale duplications only once, 100 distinct "canonical" ISH elements can be defined for the two strains. Of these, 79 are located in regions that match to each other and were analyzed with respect to ISH element mobility. Positional analysis shows that six types of ISH elements possess a high transposition frequency (group M, mobile: ISH1, ISH2, ISH3, ISH4, ISH8, and ISH11). The following description concentrates on the 65 individual elements that belong to this ISH subset.

As shown in Fig. 4, a total of 65 copies of group M elements are present either in both strains or in only one of the two strains. Of the 65 group M elements, 20 copies are located at positions conserved in both genomes, whereas 16 copies are specific for strain R1 and 29 are specific for strain NRC-1. Thus, only 31% of the mobile ISH elements are found in analogous positions, proving the very high mobility of group M ISH elements in *Halobacterium*.

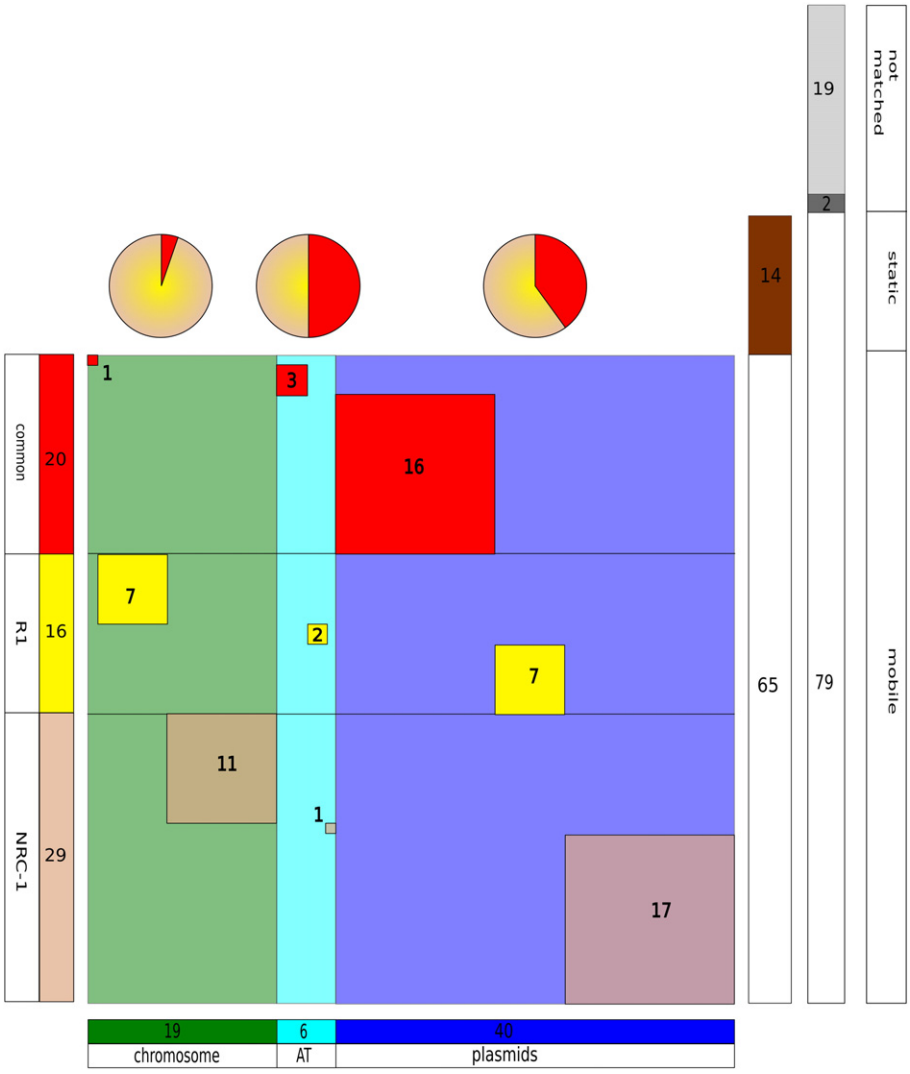

Fig. 4. Distribution of ISH elements among matching genome regions. The boxes to the right indicate grouping of the "canonical" ISH elements from strains R1 and NRC-1. When large-scale duplications are considered only once, there are a total of 100 ISH elements. Of these, 21 occur in unmatched regions restricted to strain R1 (light gray) or NRC-1 (dark gray). Of the ISH elements found in matching regions, 14 belong to the static group S (brown) and 65 belong to the mobile group M. Only one-third of the 65 mobile ISH elements in the matching genome regions occur at analogous positions (red), the remainder are specific for either strain R1 (bright yellow) or strain NRC-1 (light brown). The box at the bottom indicates the occurrence of ISH elements in genome sections. The GC-rich parts of the chromosome (green), the 60-kb AT-rich island (light blue, marked AT), and the plasmids (dark blue) are shown. The AT-rich island is separated from the remainder of the chromosome as its ISH element density is 10 times higher and similar to that of the plasmids. The central square area indicates common and strain-specific mobile ISH elements for the three genome sections. As illustrated by the pie charts, about half of the ISH elements are present in analogous positions on the plasmids and the 60-kb AT-rich island. In contrast, the GC-rich part of the chromosome contains only a single analogous mobile ISH elements. The remaining mobile ISH elements, 95%, are strain-specific.

Furthermore, a distinct bias is observed with respect to the genomic localization of strain-specific mobile ISH elements compared to those in an analogous position (Fig. 4). ISH elements are highly overrepresented (ca. 1 per 10 kb) in plasmids and in the AT-rich island of the chromosome [32], which probably originated from a plasmid integration event. The remainder of the chromosome shows a 10-fold lower density of ISH elements (only ca. 1 per 100 kb). About half of the mobile ISH elements are located at analogous positions on the plasmids and the AT-rich island. In sharp contrast, all except one (i.e., 18 of 19) of the chromosomal ISH elements show a strain-specific location (Figs. 4 and 5). From this, we conclude that the chromosomal DNA of the common ancestor of strains R1 and NRC-1 was virtually free of ISH elements. We also want to emphasize that strain-specific ISH elements outnumber single-base sequence differences and insertion/deletion events, especially in the chromosome.

Halobacterium salinarum strains R1 and NRC-1 originate from the same natural isolate and diverged by evolution in the laboratory

From the analysis provided above, it is evident that strains R1 and NRC-1 belong to the same species, *Hbt. salinarum*. The assignment of "*Halobacterium* sp. NRC-1" [7] to the species of *Hbt. salinarum* has already been suggested on the basis of taxonomic analysis [9]. The sequence comparison verifies this assignment.

In addition, based on three lines of evidence, we conclude that strains R1 and NRC-1 do not represent independent strains but most likely originate from the same cultivation event of a natural isolate. In this scenario, all differences between the two strains originate from evolution in the laboratory. First, the two chromosomes align perfectly and show only 12 distinct sequence differences when ISH element-related variations are excluded. Second, the extremely high sequence conservation is also found for the plasmid sequences, despite the major incompatibility of the overall plasmid architectures. Third, outside the AT-rich region, the chromosome of the ancestor (the initially cultivated isolate) was virtually free of ISH elements as deduced from the minimal number of ISH elements that occur in analogous positions in the two strains. With one exception, such colocalized ISH elements are restricted to the AT-rich island, which is considered to represent a plasmid integration event. The additional chromosomal copies of the mobile ISH elements were probably acquired due to the reduced selection pressure typical of cultivation in the laboratory. Alternatively, both strains could have been affected by transposition bursts, which have been described to occur in *Halobacterium* upon cell storage at 4 °C for long times and may also occur when cells are subjected to other stress factors [33].

Evolution in the laboratory may also have caused rearrangements of the plasmids. However, based on five observations, we consider it unlikely that both of the reported overall plasmid architectures can be simultaneously correct. While the overall plasmid architecture reported here for strain R1 has been validated by additional independent methods, especially cosmid analysis, no such evidence is to the best of our knowledge available for the plasmids from strain NRC-1. Taken together, we consider it unlikely that the reported architecture of the plasmids from strain NRC-1 is correct.

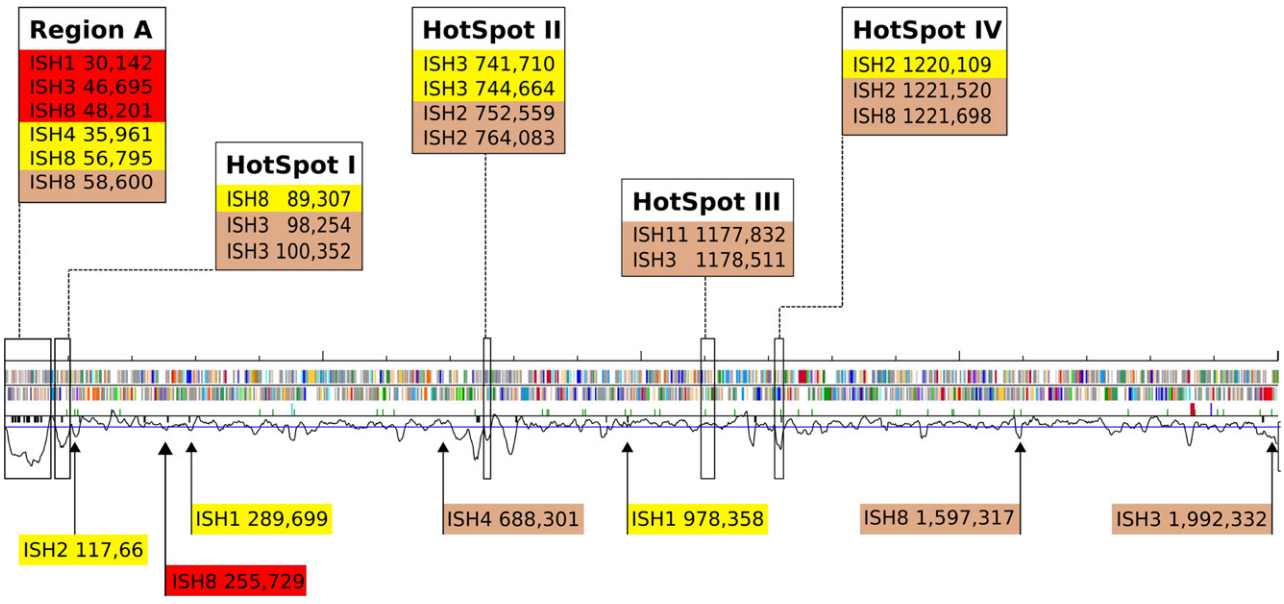

Fig. 5. Localization of mobile ISH elements in the halobacterial chromosome. Schematic representation of the *Hbt. salinarum* chromosome and the position of the "mobile" group M ISH elements. The relative GC content of the chromosome is shown (bottom line), as are the genes for stable RNAs (ticks above the central line) as well as insertion elements and other transposase-coding genes (ticks below the central line). Scaling bars are given every 100 kb above the top line. Boxes over the chromosome indicate the localization of the AT-rich island (region A) and the four chromosomal hot spots of insertion (hot spots I to IV). Dispersed ISH elements are depicted below. ISH elements that occur in analogous positions in the two halobacterial strains are indicated by red boxes, elements specific for strain R1 by yellow boxes, and those specific for strain NRC-1 by light brown boxes. The ISH type and chromosomal positions are specified for each individual copy.

First, the high stability of the plasmids at the DNA sequence level is incompatible with a severe instability at the genome structure level. In contrast, high stability of the chromosome sequences matches the high stability at the genome structure level (complete colinearity over 1.9 Mb). Second, large-scale duplications beyond 30 kb are an extreme challenge for genome assembly, which thus should be considered preliminary unless supported by additional independent methods. Third, all colinearity breakpoints among plasmids from strains R1 and NRC-1 are associated with ISH elements. In our hands, ISH elements frequently resulted in misassemblies by the applied genome assembly program. The same program was also used to assemble the NRC-1 genome. Several of the breakpoint-associated ISH elements contain target duplications in strain R1 but not in strain NRC-1. Fourth, the plasmids from strain NRC-1 contain the same 350 kb of nonduplicated sequence that is present in the plasmids from strain R1. It seems unlikely that these 350 kb exist as four plasmids in one strain but as two in the other, especially when matching regions are shuffled between the plasmids in a nontrivial way. Fifth, it seems unlikely that the plasmids in one strain contain extensive inverted duplications while the plasmids in the other strain do not.

Thus, although evolution in the laboratory may also have caused rearrangements of the plasmids, a detailed analysis of plasmid structure evolution has to be postponed. The validation or correction of the plasmids from strain NRC-1 is considered a prerequisite for such an analysis to ensure that all differences reflect biological variation.

The claim that strains R1 and NRC-1 are likely to originate from the same cultivation event is consistent with the conclusions of Grant [34], who attempted to trace the origin of *Halobacterium* strains and concluded that “in all probabilities, NRL and NRC-1 are one and the same and are held as *H. halobium* DSM 670.” Strain R1 (DSM 671) is a spontaneous gas-vesicle-free mutant of DSM 670. Consistent with this, there is a strain-specific ISH element in the promoter of the p-vac region on pHS1. Other strain-specific ISH elements affect protein-coding genes. These include restriction/modification systems (one in each strain), two TATA-binding proteins (*tfbB* and *tfbF* in strain R1), and a gene that is considered to be involved in folate biosynthesis in strain R1 (*pabA*) and a CDC6 protein homolog in strain NRC-1.

From genome sequences to the biology of Halobacterium salinarum

The genome sequences verified that haloarchaea have a high number of the basal transcription factors (TATA box-binding proteins and TFB), which was first noticed in *Haloferax volcanii* [35]. The genome sequence also revealed an unusually high number of copies for other genes such as *ftsZ* or *cdc6*. It was already known that *Hbt. salinarum* harbors several different transducers involved in phototaxis and chemotaxis [36,37]. However, only the genome sequence has shown the full spectrum of 18 genes for transducer proteins that are present in *Hbt. salinarum,* disclosing its high potential to react to a variety of physical and chemical environmental stimuli.

The availability of the genome sequences has allowed a variety of new studies that have led to a deeper understanding of the biology of *Hbt. salinarum* and makes it one of the best-studied archaeal model species [1,2]. Here we mention the results of a few studies that have been performed with strain R1. A thorough proteome analysis has been performed that led to the experimental detection of a high fraction of the cytosolic proteome [17]. This was complemented by the characterization of the membrane proteome [27] and by quantitative evaluation of membrane proteome differences in cells grown aerobically and anaerobically [24]. In addition, the low-molecular-weight proteome has recently been characterized [26], revealing the occurrence of many very small proteins, which had been systematically overlooked when using standard proteomic techniques. Therefore, a whole class of proteins was discovered that has not been studied in *Hbt. salinarum* or any other species yet. It includes many proteins with DNA/RNA binding domains that might be involved in regulatory processes [38]. A further study concentrated on the N-terminal maturation of proteins and revealed that, in contrast to bacterial proteins, a considerable fraction of haloarchaeal proteins is modified by posttranslational N-terminal acetylation [25].

Transcriptome analysis with a whole-genome DNA microarray has also been established and three studies shall exemplify different applications. (i) The microarray was used to compare transcription profiles of aerobically and phototrophically grown cells [39]. (ii) The microarray was used to study cell-cycle-specific transcript level oscillations [40]. It was found that the fraction of regulated genes is much smaller than that in the few model species studied previously. (iii) The microarray was used to compare the fractions of free and polysome-bound transcripts for all genes. For a considerable portion of the genes, differential translational control was discovered in the exponential compared to the stationary growth phase [41].

In addition to the functional genomic approaches aimed at a global overview of biological processes, several studies concentrating on specific proteins or aspects have been based on the availability of the genome sequence. It was discovered that lipid-anchored substrate-binding proteins act as sensors for amino acids and osmoprotectants in the quasi-periplasmic space and are crucial for chemotaxis toward these substrates [42]. Also, a membrane potential sensor that allows the cell to approach environments suitable for high membrane energization was identified [43]. Two novel protein families were discovered that interact and form complexes with prokaryotic structural maintenance of chromosomes (SMC) proteins, which are essential for chromosome segregation [44]. Up until then, prokaryotic SMC proteins were thought to act as simple homodimers. Furthermore, the genome sequence and especially the faithful reconstruction of the plasmid architecture allowed the quantitation of the copy number of the four replicons. Surprisingly, it was discovered that *Hbt. salinarum* is highly polyploid and that the chromosome copy number is considerably higher than the plasmid copy number [45]. Moreover, the systematic determination of the 5′ and 3′ ends of haloarchaeal transcripts has revealed that

a much higher fraction of transcripts is leaderless than had been predicted in silico [46].

Taken together, the availability of the genome sequences of the two laboratory strains of *Hbt. salinarum* has led to many important results in diverse areas of haloarchaeal biology, including genome copy number control, initiation of transcription and translation, transcriptional and translational regulation of gene expression, posttranslational modification, sensing environmental stimuli, and metabolic adaptation to various environmental conditions.

Summary and conclusions

The genome of *Hbt. salinarum* strain R1 consists of a major chromosome and four megaplasmids. Extensive cosmid analysis was used to validate the structure of the plasmids since large-scale duplications and the ample presence of ISH elements severely interfere with genome assembly procedures.

A high-quality protein-coding gene set has been obtained by rigorous evaluation of automatic gene finder data, which are highly error prone for GC-rich genomes. More than 68% of the resulting gene set has been confirmed by stringently evaluated proteomic data. These include 606 proteins for which the N-terminal peptide has been reliably identified, thus validating the assigned start codon. Large parts of the gene set are furthermore supported by homology data from intergenomic comparison to other halophiles. The HaloLex genome annotation and visualization system, which allows the integration of genomic, proteomic, sequence homology, and various other types of data, was of fundamental importance for the continuous improvement of the protein-coding gene set.

Comparison to the sequence of strain NRC-1 shows complete colinearity and an astonishingly small number of sequence differences, proving the high quality of the raw genome sequence data. The plasmids from strain R1 contain 210 kb of sequence that is not present in strain NRC-1. The remaining 350 kb match to each other and are identical except for one base change and one hot spot of sequence differences. In contrast, there are major differences in the number and overall structure of the plasmids. The pattern of duplications reported for strain NRC-1 is inconsistent with that observed for strain R1. Specifically, we could not find any evidence for inverted duplications.

Despite this near identity of the DNA sequences of strains R1 and NCR-1, there are major differences in the protein-coding gene set that affect 20% of the genes. Most of the differences are due to alternative start codon selection.

From the strict colinearity of the chromosomes and the virtual sequence identity at the DNA level for both the chromosome and the plasmids, we conclude that *Hbt. salinarum* strains R1 and NRC-1 originate from the same cultivation event of a natural isolate. Accordingly, all current differences have been acquired during cultivation in the laboratory and are due to events that occurred over the past few decades. Consistent with this, the majority of the differences is due to transposition of mobile ISH elements. Genome-scale analysis of analogous and strain-specific ISH elements indicates that the chromosome of the natural isolate was virtually free of ISH elements.

Materials and methods

Genome sequencing and assembly

Halobacterium salinarum strain R1 (DSM 671) was sequenced with 9.6-fold sequence coverage using a shotgun clone library (average insert size of 1.4 kb) and assembled with the Phred–Phrap–Consed package [47]. Major genome assembly problems were encountered due to the large number of ISH elements and large-scale duplications. These were resolved by cosmid and PCR analysis. Cosmid end sequences were positioned onto the assembled contigs using BLAST [48] and analyzed using PERL scripts. Cosmids that bridge distinct contigs were used to trigger sequencing on cosmids or on PCR products, which were generated either from cosmids or directly from the genome. We used a "mini-assembly strategy" (detailed in Supplemental Text S8) to guide the assembly program Phrap. The final sequences of the chromosome and plasmids were assembled separately. Large-scale duplications were computed within the context of pHS1 and carefully inspected to exclude undetected polymorphisms. The resulting sequences were then used for the duplicated regions of plasmids pHS2 and pHS4, respectively. The points of ring opening were chosen to allow for maximum consistency with the sequences reported for strain NRC-1.

Gene prediction

Initial gene prediction was performed by D. Frishman (MIPS) using the ORPHEUS program [49] and immediately showed a very severe ORF overprediction problem. Since then, the gene set has been continually improved, in particular through (a) evaluation of gene context as well as characteristic halophilic p*I* and amino acid distribution patterns using the HaloLex genome annotation tools, (b) comparison of the ORF set to that from strain NRC-1 [7], (c) correlation with emerging genome-wide proteomic data for *Hbt. salinarum* [17,25], (d) extension of the ORF set to include all six-frame translations with at least 100 codons, and (e) comparison to the ORF sets from *Nmn. pharaonis* [10] and *Hqr. walsbyi* [15].

Acknowledgments

We thank Bettina Brustmann for expert technical assistance, Ann-Kathrin Werenskiold and Kathrin Klee for major support during the preparation of the manuscript, Jan Wolfertz and Volker Hickmann for bioinformatics support, and Dimitrij Frishman and the MIPS annotation team for the initial ORF prediction.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.01.001.

References

[1] J. Soppa, From genomes to function: haloarchaea as model organisms, Microbiology 152 (2006) 585–590.
[2] S. Dassarma, B.R. Berquist, J.A. Coker, P. Dassarma, J.A. Muller, Post-genomics of the model haloarchaeon Halobacterium sp. NRC-1, Saline Syst. 2 (2006) 3.
[3] A. Ventosa, A. Oren, Halobacterium salinarum nom. corrig., a name to replace Halobacterium salinarum (Elazari-Volcani) and to include Halobacterium halobium and Halobacterium cutirubrum, Int. J. Syst. Bacteriol. 46 (1996) 347.
[4] F. Pfeifer, G. Weidinger, W. Goebel, Genetic variability in Halobacterium halobium, J. Bacteriol. 145 (1981) 375–381.
[5] F. Pfeifer, U. Blaseio, Insertion elements and deletion formation in a halophilic archaebacterium, J. Bacteriol. 171 (1989) 5135–5140.

[6] K. Brugger, et al., Mobile elements in archaeal genomes, FEMS Microbiol. Lett. 206 (2002) 131–141.
[7] W.V. Ng, et al., Genome sequence of Halobacterium species NRC-1, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 12176–12181.
[8] R.A. Alm, T.J. Trust, Analysis of the genetic diversity of Helicobacter pylori: the tale of two genomes, J. Mol. Med. 77 (1999) 834–846.
[9] C. Gruber, et al., Halobacterium noricense sp. nov., an archaeal isolate from a bore core of an alpine Permian salt deposit, classification of Halobacterium sp. NRC-1 as a strain of H. salinarum and emended description of H. salinarum, Extremophiles 8 (2004) 431–439.
[10] M. Falb, et al., Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis, Genome Res. 15 (2005) 1336–1343.
[11] M. Aivaliotis, et al., Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis, J. Proteome Res. 6 (2007) 2195–2204.
[12] A.C. McHardy, A. Goesmann, A. Puhler, F. Meyer, Development of joint application strategies for two microbial gene finders, Bioinformatics 20 (2004) 1622–1631.
[13] P. Nielsen, A. Krogh, Large-scale prokaryotic gene prediction and comparison to genome annotation, Bioinformatics 21 (2005) 4322–4329.
[14] B.R. Berquist, S. DasSarma, An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, Halobacterium sp. strain NRC-1, J. Bacteriol. 185 (2003) 5959–5966.
[15] H. Bolhuis, et al., The genome of the square archaeon Haloquadratum walsbyi: life at the limits of water activity, BMC Genomics 7 (2006) 169.
[16] N.S. Baliga, et al., Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea, Genome Res. 14 (2004) 2221–2234.
[17] A. Tebbe, et al., Analysis of the cytosolic proteome of Halobacterium salinarum and its implication for genome annotation, Proteomics 5 (2005) 168–179.
[18] F. Pfeifer, M. Betlach, Genome organization in Halobacterium halobium: a 70 kb island of more (AT) rich DNA in the chromosome, Mol. Gen. Genet. 198 (1985) 449–455.
[19] A. Ruepp, J. Soppa, Fermentative arginine degradation in Halobacterium salinarium (formerly Halobacterium halobium): genes, gene products, and transcripts of the arcRACB gene cluster, J. Bacteriol. 178 (1996) 4942–4947.
[20] F. Pfeifer, P. Ghahraman, Plasmid pHH1 of Halobacterium salinarium: characterization of the replicon region, the gas vesicle gene cluster and insertion elements, Mol. Gen. Genet. 238 (1993) 193–200.
[21] W.L. Ng, S. DasSarma, Minimal replication origin of the 200-kilobase Halobacterium plasmid pNRC100, J. Bacteriol. 175 (1993) 4584–4596.
[22] K. Konstantinidis, et al., Genome-wide proteomics of Natronomonas pharaonis, J. Proteome Res. 6 (2007) 185–193.
[23] F. Veloso, G. Riadi, D. Aliaga, R. Lieph, D.S. Holmes, Large-scale, multigenome analysis of alternate open reading frames in bacteria and archaea, Omics 9 (2005) 91–105.
[24] B. Bisle, et al., Quantitative profiling of the membrane proteome in a halophilic archaeon, Mol. Cell. Proteomics 5 (2006) 1543–1558.
[25] M. Falb, et al., Archaeal N-terminal protein maturation commonly involves N-terminal acetylation: a large-scale proteomics survey, J. Mol. Biol. 362 (2006) 915–924.
[26] C. Klein, et al., The low molecular weight proteome of Halobacterium salinarum, J. Proteome Res. 6 (2007) 1510–1518.
[27] C. Klein, et al., The membrane proteome of Halobacterium salinarum, Proteomics 5 (2005) 180–197.
[28] S. Mattar, M. Engelhard, Cytochrome ba3 from Natronobacterium pharaonis—an archaeal four-subunit cytochrome-c-type oxidase, Eur. J. Biochem. 250 (1997) 332–341.
[29] A.S. Mankin, N.L. Teterina, P.M. Rubtsov, L.A. Baratova, V.K. Kagramanova, Putative promoter region of rRNA operon from archaebacterium Halobacterium halobium, Nucleic Acids Res. 12 (1984) 6537–6546.
[30] W.V. Ng, et al., Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? Genome Res. 8 (1998) 1131–1141.
[31] K.F. Storch, J. Rudolph, D. Oesterhelt, Car: a cytoplasmic sensor responsible for arginine chemotaxis in the archaeon Halobacterium salinarum, EMBO J. 18 (1999) 1146–1158.
[32] F. Pfeifer, Insertion elements and genome organization of Halobacterium-halobium, Syst. Appl. Microbiol. 7 (1986) 36–40.
[33] F. Pfeifer, U. Blaseio, Transposition burst of the ISH27 insertion element family in Halobacterium halobium, Nucleic Acids Res. 18 (1990) 6921–6925.
[34] W.D. Grant, Genus I: Halobacterium Elazari-Volcani 1957, 207AL emend. Larsen and Grant 1989, 2222, in: D.R. Boone, R.W. Castenholz, G.M. Garrity (Eds.), Bergey's Manual of Systematic Bacteriology, 2nd ed., Vol. 1, Springer-Verlag, Berlin, 2001, pp. 301–305.
[35] J.N. Reeve, K. Sandman, C.J. Daniels, Archaeal histones, nucleosomes, and transcription initiation, Cell 89 (1997) 999–1002.
[36] J. Rudolph, et al., A family of halobacterial transducer proteins, FEMS Microbiol. Lett. 139 (1996) 161–168.
[37] W. Zhang, A. Brooun, J. McCandless, P. Banda, M. Alam, Signal transduction in the archaeon Halobacterium salinarium is processed through three subfamilies of 13 soluble and membrane-bound transducer proteins, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 4649–4654.
[38] V.Y. Tarasov, H. Besir, R. Schwaiger, K. Klee, K. Furtwängler, F. Pfeiffer, D. Oesterhelt, A small protein from the *bop-brp* intergenic region of *Halobacterium salinarum* contains a zinc finger motif and regulates *bop* and *crtB1* transcription, Mol. Microbiol. 67 (2008) 772–780.
[39] J. Twellmeyer, et al., Microarray analysis in the archaeon Halobacterium salinarum strain R1, PLoS ONE 2 (2007) e1064.
[40] A. Baumann, C. Lange, J. Soppa, Transcriptome changes and cAMP oscillations in an archaeal cell cycle, BMC Cell Biol. 8 (2007) 21.
[41] C. Lange, A. Zaigler, M. Hammelmann, J. Twellmeyer, G. Raddatz, S.C. Schuster, D. Oesterhelt, J. Soppa, Genome-wide analysis of growth phase-dependent translational and transcriptional regulation in halophilic archaea, BMC Genomics 8 (2007) 415.
[42] M.V. Kokoeva, K.F. Storch, C. Klein, D. Oesterhelt, A novel mode of sensory transduction in archaea: binding protein-mediated chemotaxis towards osmoprotectants and amino acids, EMBO J. 21 (2002) 2312–2322.
[43] M.K. Koch, D. Oesterhelt, MpcT is the transducer for membrane potential changes in Halobacterium salinarum, Mol. Microbiol. 55 (2005) 1681–1694.
[44] J. Soppa, et al., Discovery of two novel families of proteins that are proposed to interact with prokaryotic SMC proteins, and characterization of the Bacillus subtilis family members ScpA and ScpB, Mol. Microbiol. 45 (2002) 59–71.
[45] S. Breuert, T. Allers, G. Spohn, J. Soppa, Regulated polyploidy in halophilic archaea, PLoS ONE 1 (2006) e92.
[46] M. Brenneis, O. Hering, C. Lange, J. Soppa, Experimental characterization of cis-acting elements important for translation and transcription in halophilic archaea, PLoS Genetics 3 (2007) e229.
[47] D. Gordon, C. Abajian, P. Green, Consed: a graphical tool for sequence finishing, Genome Res. 8 (1998) 195–202.
[48] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[49] D. Frishman, A. Mironov, H.W. Mewes, M. Gelfand, Combining diverse evidence for gene recognition in completely sequenced bacterial genomes, Nucleic Acids Res. 26 (1998) 2941–2947.